Piloting access to the Belgian web archive for scientific research: a methodological exploration

Sally Chambers, Eveline Vlassenroot, Peter Mechant and Friedel Geeraert

Engaging with Web Archives: 'Opportunities, Challenges and Potentialities' #EWAVirtual
21-22 September 2020, Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland.

KBR

promise

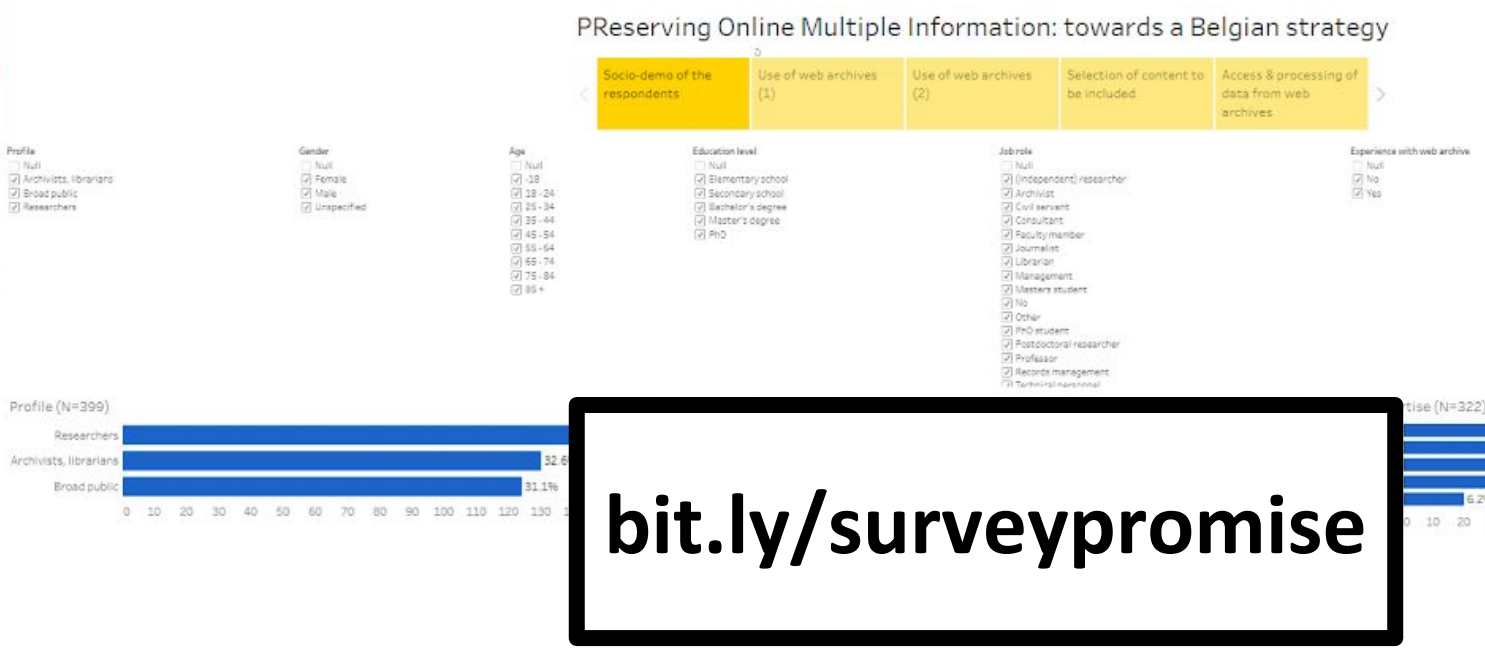
GHENT UNIVERSITY

Exploring how to pilot access to the Belgian web archive for scientific research?

- How do existing web-archives provide access to their collection for research?
- What can we learn from existing research initiatives using web archives?
- Using personas to evaluate software for the *discovery*, *delivery* & *analysis* of archived web data

www.kbr.be/en/projects/promise-project/

User requirements for a web archive



bit.ly/surveypromise

PROMISE (PReserving Online Multiple Information: towards a Belgian StratEgy)

More information

KBR

RSITY

Reviewing how existing web archives provide access to their collections for research

Eveline Vlassenroot – imec-MICT Ghent University

KBR

promise

GHENT UNIVERSITY

Different types of access (Winters, 2019)

1) Read an archived web page
2) Download and manipulate data
3) Availability of appropriate archival frameworks and tools to facilitate navigation and research
4) Access facilitated by publication of research findings and data

Winters, Jane (2019). *Negotiating the archives of UK web space.* In: The Historical Web and Digital Humanities: the Case of National Web Domains, ed. Niels Brügger and Ditte Laursen. Routledge, Abingdon.

Different types of access

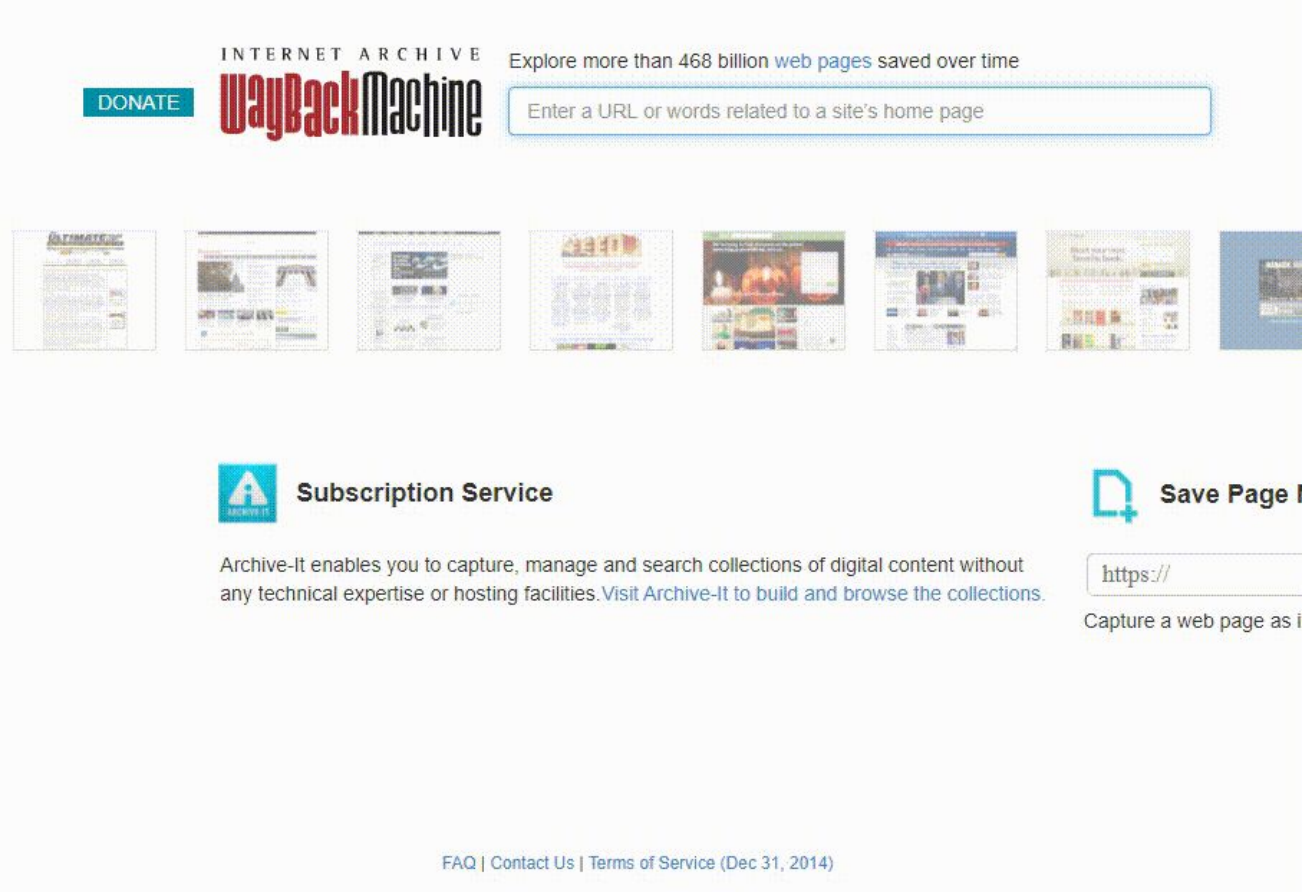
1) Read an archived web page
2)
3)
4)

Country	*Initiative*	Web archive is available for consulta-tion
Canada	*Library and Archives Canada (LAC)*	✓
Canada	*Bibliothèque et Archives nationales du Québec (BAnQ)*	some
Denmark	*Netarkivet*	✓
Estonia	*Eesti Veebiarhiiv*	some
France	*Bibliothèque nationale de France (BnF)*	✓
France	*Institut national de l'audiovisuel (INA)*	✓
Hungary	*National Széchényi Library*	x
Ireland	*National Library of Ireland*	✓
Luxembourg	*Bibliothèque nationale du Luxembourg (BnL)*	✓
New-Zealand	*National Library of New Zealand (NLNZ)*	✓
Switzerland	*Webarchiv Schweiz*	✓
The Netherlands	*KB Webarchief (KB)*	✓
The Netherlands	*Nationaal Archief (NA)*	some
UK	*UK Web Archive (UKWA)*	✓
USA	*George Washington University Libraries (GWUL)*	✓

KBR

Different types of access

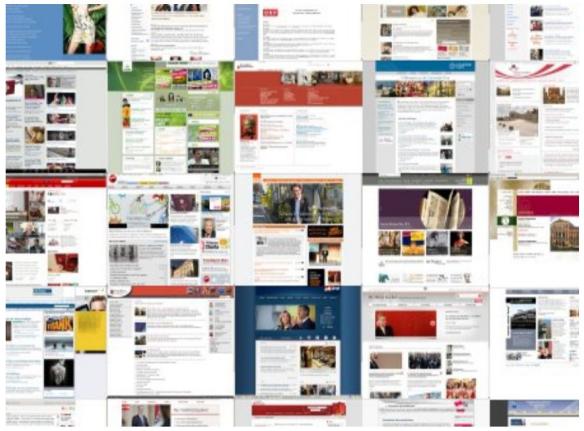
1)
2) Download and manipulate data
3)
4)



KBR

GHENT UNIVERSITY

Web archives as datasets



Webarchive Austria

Metadata for 2M archived webpages starting 2009

[Web Archive Austria]

KB LAB

Web Collection Chinese Netherlands

This special web collection contains a selection of archived websites from the Chinese community in the

[Web Collection: Chinese Netherlands]

KB LAB

Web collection internet archaeology Euronet-Internet (1994-2017)

The special web collection internet archaeology Euronet-Internet (1994-2017) [Webcollectie internetarcheologie Euronet: websites gehost bij provider Euronet-Internet

[Web Collection: Internet Archaeology]

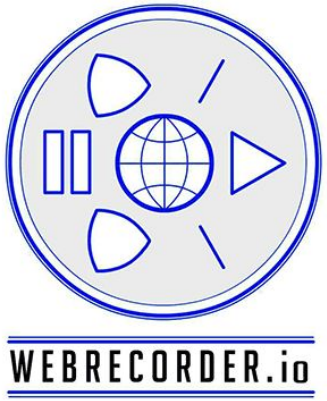
KBR

promise

GHENT UNIVERSITY

Different types of access

1)
2)
3) Availability of appropriate archival frameworks and tools to facilitate navigation and research
4)



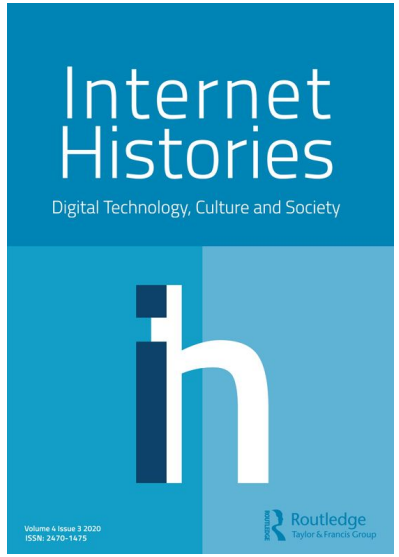
KBR

promise

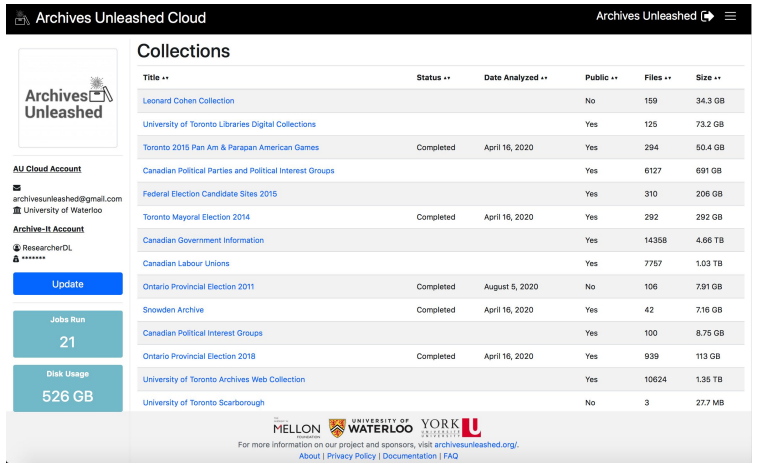
GHENT UNIVERSITY

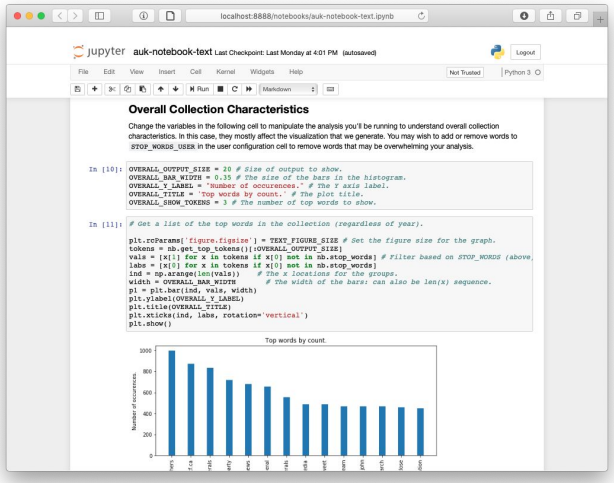
Different types of access

1)
2)
3)
4) Access facilitated by publication of research findings and data



[Archives Unleashed Notebooks]





[Web Archive Datasets in Zenodo & Dataverse]

KBR

promise

GHENT UNIVERSITY

What can we learn from existing research initiatives using web archives?

Sally Chambers – Ghent Centre for Digital Humanities, Ghent University

KBR

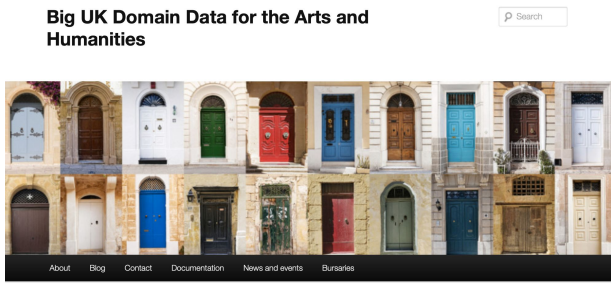
promise

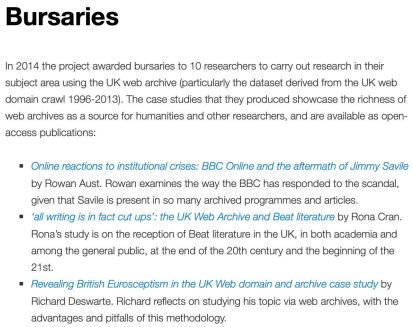
GHENT UNIVERSITY

Research-use of web archives

What can we learn from existing initiatives?

- **Awareness raising:** web archives as a research resource





Bursaries

In 2014 the project awarded bursaries to 10 researchers to carry out research in their subject area using the UK web archive (particularly the dataset derived from the UK web domain crawl 1996-2013). The case studies that they produced showcase the richness of web archives as a source for humanities and other researchers, and are available as open-access publications:

- *Online reactions to institutional crises: BBC Online and the aftermath of Jimmy Savile* by Rowan Aust. Rowan examines the way the BBC has responded to the scandal, given that Savile is present in so many archived programmes and articles.
- *'all writing is in fact cut ups': the UK Web Archive and Beat literature* by Rona Cran. Rona's study is on the reception of Beat literature in the UK, in both academia and among the general public, at the end of the 20th century and the beginning of the 21st.
- *Revealing British Eurosceptism in the UK Web domain and archive case study* by Richard Deswarte. Richard reflects on studying his topic via web archives, with the advantages and pitfalls of this methodology.

- **Community building:** web archives are complex datasets: we don't need to go it alone

KBR

promise

GHENT UNIVERSITY

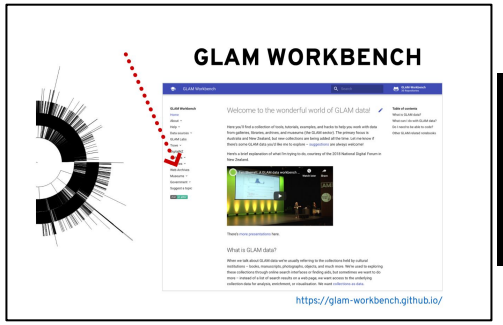
Research-use of web archives

What can we learn from existing initiatives?

- **Research collaboration:** Research network focusing on Web Archives

WARCnet

- **Research infrastructure for the archived web**: sharing data, tools and workflows

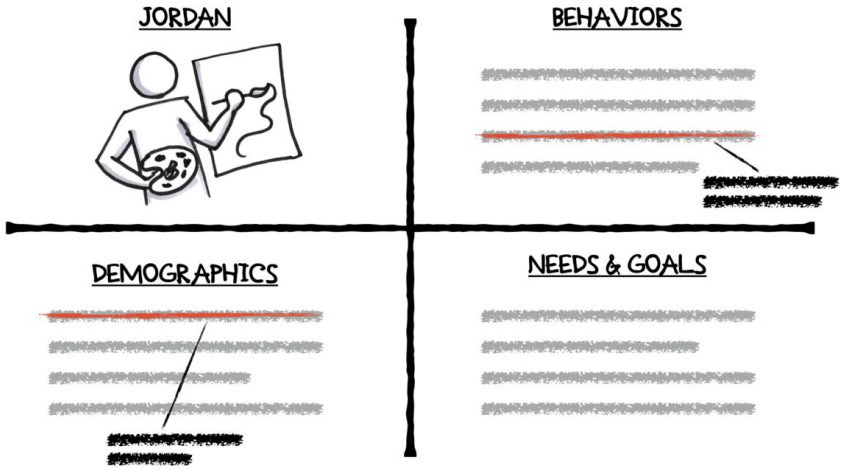


IIPC Research Speaker Series

KBR

promise

GHENT UNIVERSITY



Peter Mechant – imec-MICT, Ghent University

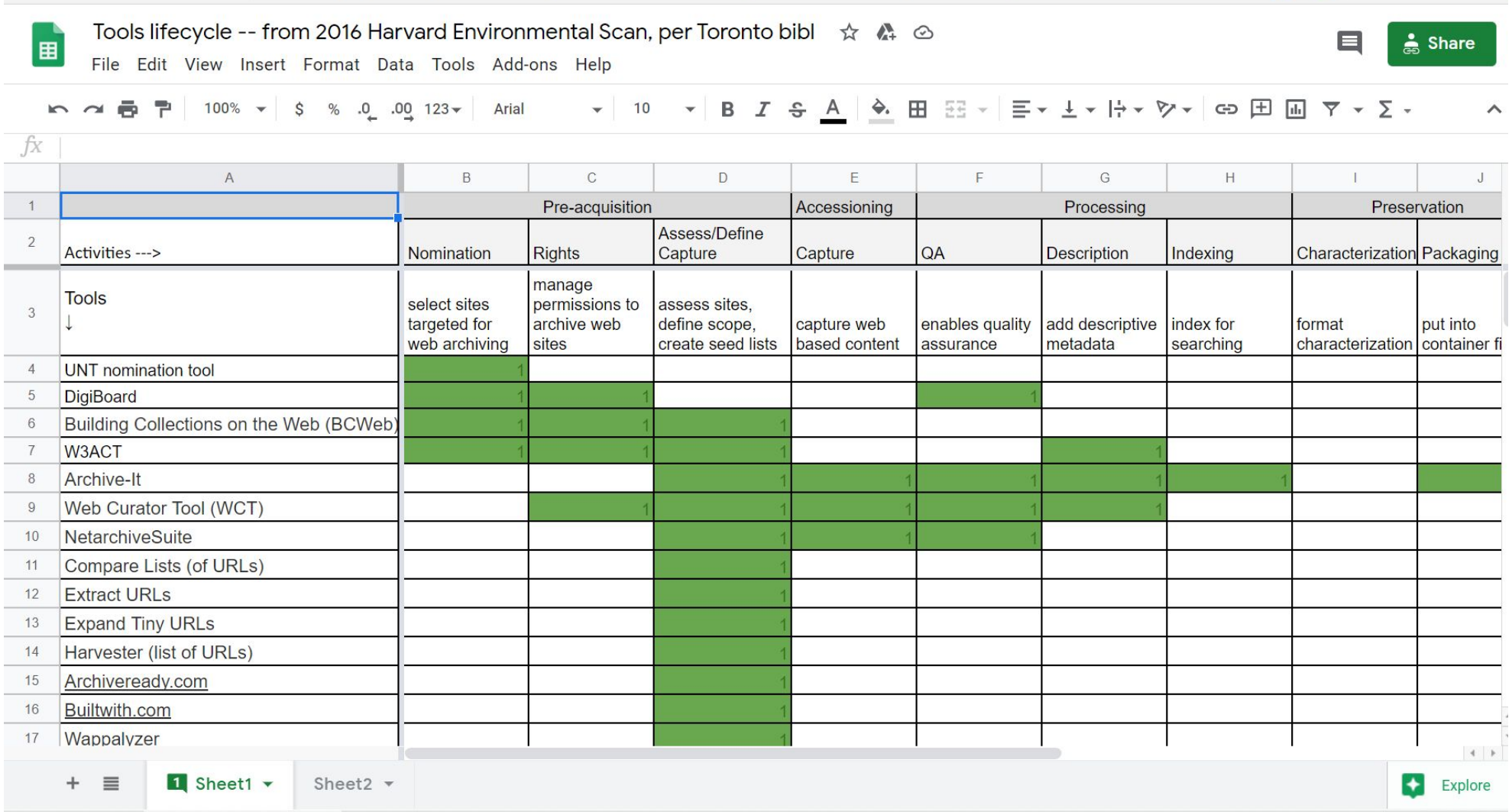
Using **personas** to evaluate software for the Discovery, Delivery and Analysis of archived web data

KBR

promise

GHENT UNIVERSITY

Step 1



Tools lifecycle -- from 2016 Harvard Environmental Scan, per Toronto bibl

File Edit View Insert Format Data Tools Add-ons Help

Share

	A	B	C	D	E	F	G	H	I	J
1		Pre-acquisition			Accessioning	Processing			Preservation	
2	Activities --->	Nomination	Rights	Assess/Define Capture	Capture	QA	Description	Indexing	Characterization	Packaging
3	Tools ↓	select sites targeted for web archiving	manage permissions to archive web sites	assess sites, define scope, create seed lists	capture web based content	enables quality assurance	add descriptive metadata	index for searching	format characterization	put into container fi
4	UNT nomination tool	1								
5	DigiBoard	1	1			1				
6	Building Collections on the Web (BCWeb)	1	1	1						
7	W3ACT	1	1	1			1			
8	Archive-It			1	1	1	1	1		
9	Web Curator Tool (WCT)		1	1	1	1	1			
10	NetarchiveSuite			1	1	1				
11	Compare Lists (of URLs)			1						
12	Extract URLs			1						
13	Expand Tiny URLs			1						
14	Harvester (list of URLs)			1						
15	Archiveready.com			1						
16	Builtwith.com			1						
17	Wappalyzer			1						

Sheet1 Sheet2 Explore

see Truman, Gail. 2016. Web Archiving Environmental Scan. Harvard Library Report - http://bit.ly/1Zok3WB https://dash.harvard.edu/bitstream/handle/1/25658314/print_HL_web_archiving_env_scan_2017.pdf?sequence=4&isAllowed=y

KBR

promise

GHENT UNIVERSITY

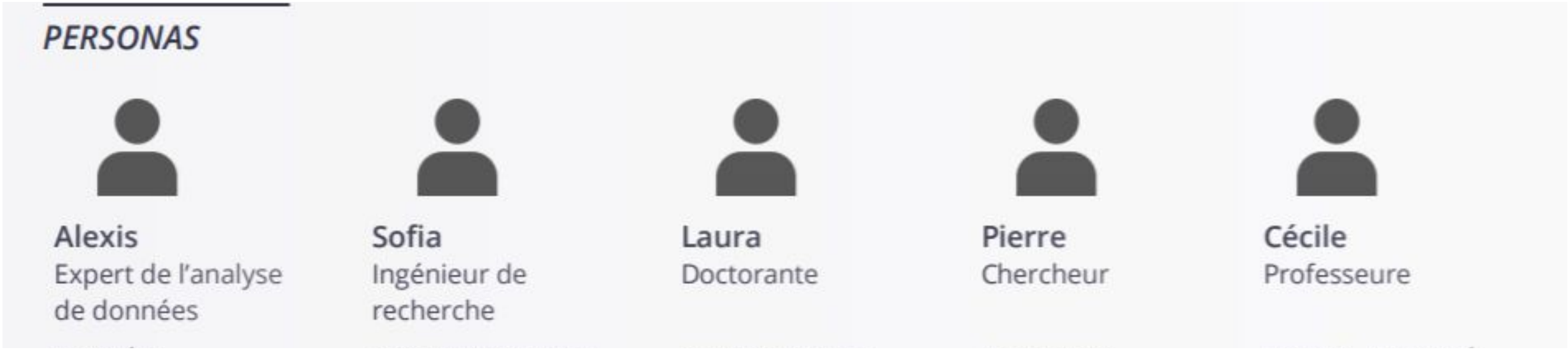
Step 2

Le projet Corpus et ses publics potentiels.

Eleonora Moiraghi

▶ **To cite this version:**

Eleonora Moiraghi. Le projet Corpus et ses publics potentiels. : Une étude prospective sur les besoins et les attentes des futurs usagers.. [Rapport de recherche] Bibliothèque nationale de France. 2018. hal-01739730



see https://hal-bnf.archives-ouvertes.fr/hal-01739730/document

Step 2

Five personas from the **CORPUS project** were considered:

Cécile, a 50-year old art history professor, with very low ICT- and software skills but with a high expertise in the domain of human and social sciences.

Sofia, a 34-year old research engineer, who helps researchers with all digital related issues and who has subsequently a rather high level of ICT- and software skills but rather low domain expertise in human and social sciences.

Laura, a 27-year old PhD-candidate in history, with mediocre ICT- and software skills and with rather high domain expertise in human and social sciences.

Alexis, a 28-year old data scientist, with very high ICT- and software skills but with a low expertise in the domain of human and social sciences. In a sense this persona can be contrasted with the persona of Cécile.

Pierre, a 36-year old researcher studying literature, who has very low ICT- and software skills but a strong expertise in the domain of human and social sciences

see https://hal-bnf.archives-ouvertes.fr/hal-01739730/document

KBR

promise

GHENT UNIVERSITY

Step 3

KBR

promise

GHENT UNIVERSITY

Step 3

	Cécile	Laura	Sofia	Pierre	Alexis
	Professor in history	PhD candidate in history	Research assistant	Researcher literature	Data scientist
Wayback Machine	+++	+++	+++	+++	+++
Archive.is	+++	+++	+++	+++	+++
Memento Time Travel	+++	+++	+++	+++	+++
Archives Unleashed Cloud	++	++	++	++	+++
Archives Unleashed Notebooks	+	++	++	+	+++
Warclight	+	++	++	+	+++
Archives Unleashed Toolkit	+	++	++	+	+++
Webrecorder pywb	+	++	++	+	+++
Browsertrix	+	++	++	+	+++
ArchiveThumbnails	+	++	++	+	+++
Shine	+	++	++	+	+++
Warcbase	+	+	++	++	+++
Natural Language Toolkit (NLTK)	+	++	+++	+	++
Leximancer	+	++	+++	+	++
WCopyfind	+	++	+++	+	+++
Mallet	+	++	+++	+	+++
CarbonDate	+++	+++	+++	+++	+++
Dorling Map Generator	+++	+++	+++	+++	+++
Raw Text to Tag Cloud Engine	+++	+++	+++	+++	+++
Gephi	+	+++	+++	+++	++

GHENT UNIVERSITY

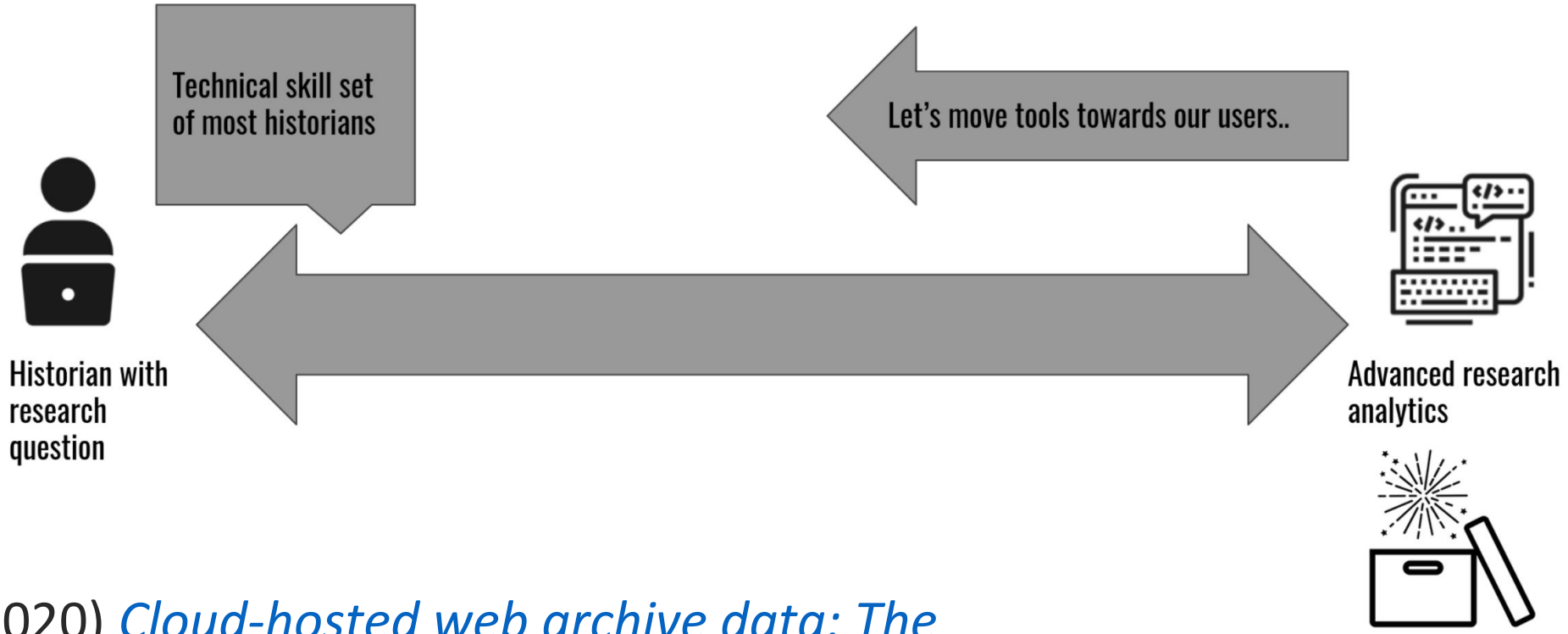
Alexis

- as data scientist does not encounter lots of problems working with the tools as he has very high ICT- and software skills.
- due to rather low expertise in the domain of human and social sciences actually using and interpreting the results of some of these tools might pose some problems.

Cécile

- possesses very low ICT- and software skills but has high domain expertise
- only the most basic tools offering simple (online) interfaces are in her reach;
- for more complex software she needs to acquire new skills or set up collaborations with other colleagues (e.g. Alexis or Sofia).

Bringing web archives closer to researchers



Ruest, N. (2020) *Cloud-hosted web archive data: The winding path to web archive collections as data*.
Archives Unleashed project

KBR

promise

GHENT UNIVERSITY

Exploring how to pilot access to the Belgian web archive for scientific research?

- Variety of access modes (including data-level access)
- Importance of collaboration with the web archiving community
- Increased digital literacy skills or interdisciplinary collaboration

www.kbr.be/en/projects/digital-research-lab/

Thank you!

Sally Chambers, Eveline Vlassenroot, Peter Mechant and Friedel Geeraert
(Sally.Chambers@UGent.be)

Engaging with Web Archives: 'Opportunities, Challenges and Potentialities' #EWAVirtual
21-22 September 2020, Maynooth University Arts and Humanities Institute, Co. Kildare, Ireland.

KBR

promise

GHENT UNIVERSITY